Lec 11 | classifiers

*classification : assign lables to objects based on object's attributes.

examples ⟶ Naive Bayesian & decision trees.

Naïve Bayesian

→ Probabilistic classifier based on Bayes' law and (naive Conditional independent assumptions.

Input variables are discrete with variations to
↳ ~~Variables are~~
algorithms that work with continuous variables as well.

output
*Probability score : Proportional to true probability
* class label : based on highest probability score.

Bayesian - use cases

*Preferred for many text classification problems.

-* use cases
  ↳ spam filtering
  ↳ Fraud detection.

ـــ فيه مثال محلول لل (classifier) هيتنزل فيما بعد
في ملف ال (examples)

* Baye's' law

$$P(c|A) = \frac{P(A \cap c)}{P(A)} = \frac{P(A|c)\,P(c)}{P(A)}$$

Naïve Bayesian classifier

Reasons to choose	Cautions
Handles missing values quite well	Numeric variables have to be discrete intervals.
Robust to irrelevent variables	Sensitive to correlated variables "Double-counting"
Easy to implement	Not good for estimating Probabilities stick to class labels
Easy to predict data	
Resistant to over-fitting	

* What gives the naïve Bayesian the advantage of being computationally inexpensive?

→ It is computationally efficient.

↳ handles very high dimensional problems.

↳ " categorical variables with lot of levels.

Lec 12

Decision tree classifier

input variables can be continuous or discrete

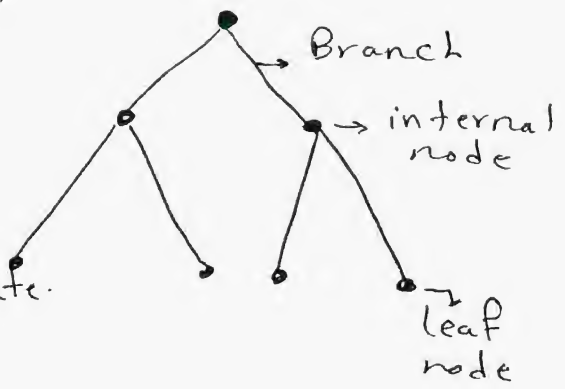
output

↳ tree describes the decision flow.

↳ trees can be converted to set of decision rules.

Branches → outcome of decision

internal node → the decision or test points. each refers to single variables or attribute.



* Decision tree classifier use cases

[1] when if-then conditions are preferred to linear models
- a) customer segmentation.
- b) financial decisions.
- c) fraud detection.

[2] when series of questions (Yes/No) are answered to arrive at classification.
- Biological species classification.

---

* week learners
- It is components (short decision trees)
- set of predictive models which will all vote and we take decisions based on combination of votes (used in ensemble techniques)

Decision tree classifier

Reasons to choose	Cautions.
a) Easy to score data. b) Computationally efficient to build. c) handles variables that have non-linear effect on outcome. d) decision rules are easy to understand. (in Principle)	a) decision surfaces can only be axis-aligned. b) related to over-fit problem. c) in practice, decision rules can be complex. d) ~~deep~~ tree structure is sensitive to small changes in training data.

Typical questions	recommended methods.
* Do I want class probabilites rather than class labels.	Logistic regression Decision tree
* Do i want insight into how variables affect model?	Logistic & Decision tree
Is problem high-dimensional?	Naive Bayes
Do I suspect some of inputs are correlated?	Decision & Logistic
Do I suspect some of inputs are irrelevant?	Decision tree & Naive Bayes
Are there mixed variables types	Decision tree & Logistic regression.

Lec 13

Time series

↳ basic research methodology in which data for one or more variables collected for many observations at different time periods.

Key component of time-series

[1] Trend Component

↳ trend is long term movement in a time-series.
↳ underlying direction that can be +ve or -ve.

[2] Seasonal Component

↳ component of variation in time series which is dependent on time of year.
~~describes regular variation~~

[3] cyclic Component

↳ cyclical variations of non-seasonal nature, whose periodicity is unknown.

[4] Random Component

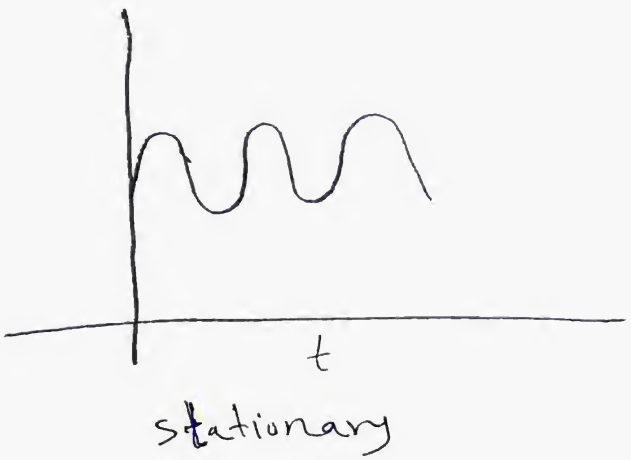
↳ Random or chaotic values left over when other components of series have been accounted for.

* use cases of time series

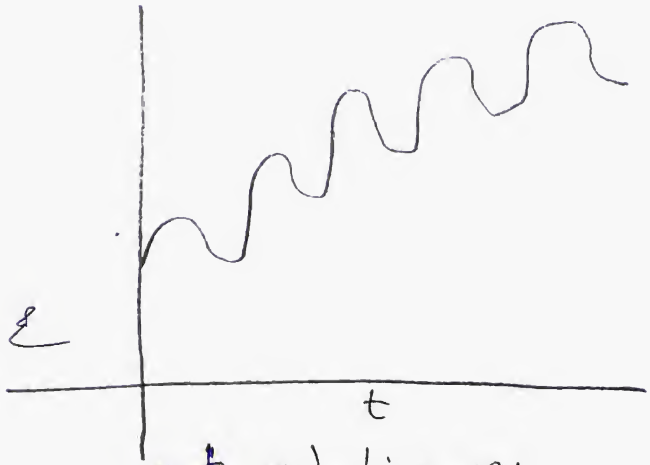
1) Economics / finance (profits, imports)
2) Environment (amount of pollutants)
3) Medicine: Blood pressure measurements.
4) Sociology: Crime rate

* stationary sequences:

↳ random sequence in which joint probability distribution does not vary over time.
↳ mean, variance and auto correlations don't change in sequence over time.



stationary



not stationary

* Constant mean & Variance

↳ mean of series shouldn't be function of time.
↳ variance " " " " " " " ".

Notes

→ Variance refers to spread of data set

→ Covariance " " measure of how two random variables will change together.

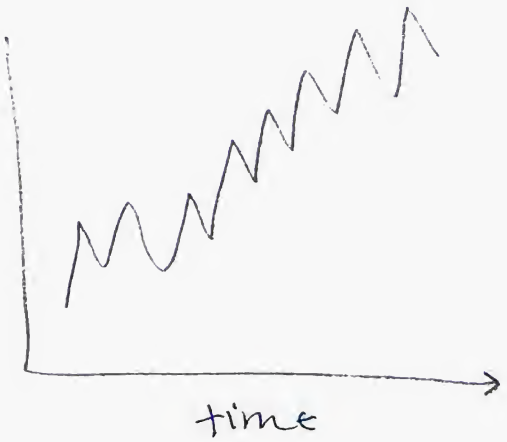
↳ used to calculate correlation between variables.

Detrending

↳ Pre-Processing step to prepare time series for analysis by methods that assume stationarity.

غير ساكن (non-stationary) ← الشكل →

(de-trend) نعمل



→ linear model

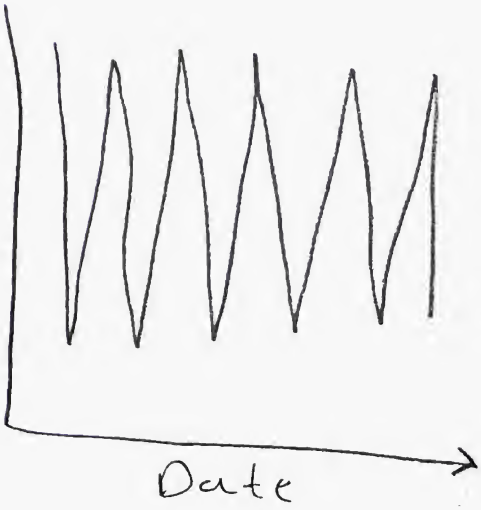
$$T_t = m \cdot t + b$$

→ detrended series

$$Y_t^1 = Y_t - T_t$$

(trend component) لـ مكون الـ

(time series) سلسلة الـ



Differencing

↳ technique used to remove non-stationarity

$$X(t) - X(t-1) = ARMA(p, q)$$

→ done by subtract observation in current period from previous one.

⇒ ARIMA Parameters

p: AR & d: I & q: MA

Where

p: no. of autoregressive terms.

d: no. of differences.

q: no. of moving average terms.

Notes

Auto Correlations

↳ numerical values that indicate how a data series is related to itself over time.

↳ measures how strongly data values at a specified number of periods apart are correlated to each other over time.

* Auto Correlation Function (ACF)

↳ Correlation of the values of time series with itself.

↳ Helps to determine the order, q of MA model.

* Partial autoCorrelation function (PACF)

↳ AutoCorrelation calculated after removing the linear dependence of previous terms.

↳ Helps to determine order, p of AR model

AR → Auto-regression

MA → moving average

* How to do a time series analysis

1) visualize time series

↳ إنك تعمل (explore) لل (data) شيء مهم في ال (model)

↳ من غير ال (exploration) مش هتعرف ال (series) تكون (stationary) ولا لأ.

2) stationarize the series

↳ بالاستخدام هنستخدم الثلاث طرق de-trending & seasonal adjustment & differencing.

3] Plot ACF/PACF to find optimal Parameters

لـ الـ Parameters (p,d,q) ممكن تيجي باستخدام ACF, PACF

لـ لو ACF, PACF بتقل ~~Significantly~~ تدريجي، معناه

إننا ضفل الـ (series) ← stationary .

4] Build ARIMA model

← هنستخدم الـ (Parameters) عشان نبني الـ (model)

5] Make Predictions

↳ لـ حالياً نقدر نعمل (Predictions).

Time series

Reasons to choose	Cautions.
Minimal data collection. ↳ don't need to input drivers	no meaningful drivers. ↳ no explanatory value. ↳ can't stress test.
designed to handle the inherent autocorrelation of lagged time series.	↳ It's an "art form" to select appropriate Parameters.
Accounts for trends and seasonality.	→ only suitable for short term Predictions.

Lec 14

text analysis

↳ Processing & representation of text for analysis and learning tasks.

* Main challenges in text analysis:.

1) High-dimensionality
   ↳ every distinct term is dimension.
   ↳
2) data is un-structured

* Inverse document Frequency (idf)

$$idf(t) = \log [N / df(t)]$$

↳ indicates the importance of:
   ↳ search (relevance)
   ↳ classification

لـ المحاضرة مختلفة تتذاكر مسار (slide) أكثر.
للي حابب يفهمها كلها.